微服务器研究现状综述*

王聪 侯锐 张立新

1 微服务器(Micro Server)的缘起和定义

数据中心的能源消耗和空间需求成为越来越重要的经济和环境问题。为了应对由此带来的挑战，服务器设计者提出微服务器的理念，即以低功耗、中等甚至低复杂度的处理器作为核心来建造服务器。最近几年全球主要的服务器和芯片厂商纷纷围绕该理念推出一系列的高能效微服务器及其发展路线图。比如戴尔（DELL）[1]、惠普（HP）[2][3]和 SeaMicro[4]等服务器厂商发布了他们的微服务器系统（DELL PowerEdge、HP ProLiant、HP Redstone、Seamicro SM10000）；英特尔（Intel）发布了 ivy bridge 架构的 Atom 处理器 Centerton 和 Xeon E3 系列；ARM Cortex A9/A15 处理器也在进军服务器行业，全面面向服务器设计的 ARMv8 架构的 X-gene 64 位服务器处理器将在不久面世。这些系统将几十到几百个高性能功耗比的处理器集中到了一台机器或一个机架里。可以预见，微服务器是解决现代计算系统能源和空间问题的一个有前景的途径，微服务器在数据中心领域将会占据一席之地并且得到广泛的应用。

大量来自于实际应用的数据表明，对于一大类数据中心的典型工作负载，尤其是一些大数据处理的应用场景，微服务器采用的中低端处理器比传统的高端处理器在计算效能、硬件成本以及计算密度等方面均具有较大的优势。

1.1 微服务器具有更好的性能功耗比

为了验证低功耗处理器的高性能功耗比，我们在基于 Xeon 和 Atom 的两个集群上测试了一个大数据处理的典型负载[5]——基于 Hadoop 的 Mastiff。Mastiff 是一个列存储的分布式大数据管理系统。具体的测试平台包括一个基于 Intel Xeon 的高端服务器集群和一个基于 Intel Atom 的低功耗服务器集群。表 1 列出了配置细节。Xeon 集群代表数据中心常用的高性能系统，而 Atom 集群代表低功耗微处理器平台。集群中有一个节点作为主节点（master node），用于元数据（metadata）的记录、任务分配和系统管理，其余 30 个节点作为从节点（slave node），用于运行主节点分配来的任务。两个集群都使用千兆交换机进行互联。

在节点数量（31 个）、工作类型——数据加载（Data load）和数据查询（Data query）、输入数据（1TB）均相同的情况下，图 1（a）对比了两种集群的功耗和性能。可以发现，在数据加载和数据查询两种情景下完成所有工作，Atom 集群消耗的能量都只是 Xeon 集群的一半。

由于在节点数目相同的情况下 Atom 完成工作所花的时间是 Xeon 集群的两倍，为了在时间上做到公平，我们将 Xeon 集群的从节点数目减少一半，从 30 降到 15。再次测试后发现，两个集群花费的时间相当，而如图 1（b）所示，Atom 集群消耗的能量仍然是 Xeon 集群的一半。由此得到结论：在本实验环境中基于 Atom 的集群性能功耗比要高于基于 Xeon 的集群。正因如此，使用低功耗、中等复杂度处理器成为当前计算机系统发展新趋势中的一个重要方面。国际上众多研究小组的观察也得到类似结论，对于一大类数据中心的典型负载，

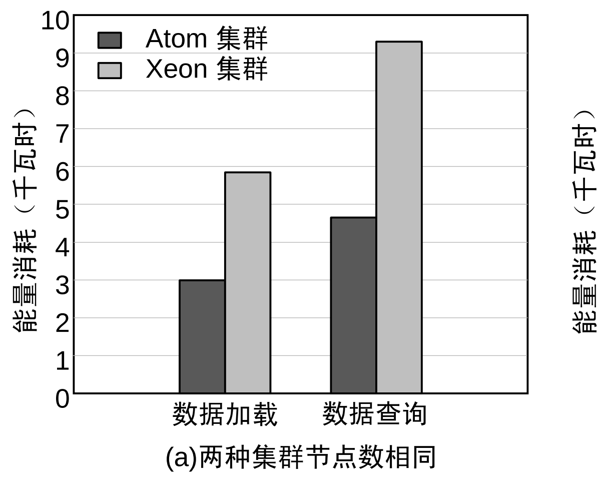
---

*本研究得到 IBM SUR 大学合作项目的资助

尤其是追求高并发，而对单线程处理能力要求不高的负载，微服务器比高端服务器具备更好的性能功耗比。

表1. 低功耗 Atom 集群和高端 Xeon 集群的配置[5]

		Atom 集群	**Xeon 集群**
节点数		31	31
机箱	型号	SuperCloud SC-R6280	Dawning I610r-H
	尺寸	2U 8 节点	1U 1 节点
节点	处理器	Intel Atom D525 (双核/4 线程/1.8 GHz /1M L2 缓存/ 13W Max TDP)	Intel xeon E5310 (4 核/4 线程/1.6 GHz /8M L2 缓存/ 80W Max TDP)
	芯片组	Intel ICH9R	Intel 5100+ICH9R
	内存	2×2GB DDR3 Non-ECC 800MHz SO-DIMM	2×2GB FBD DDR2 667MHz ECC
	磁盘	500GB SATA 5400RPM	1TB SATA 7200RPM
	网络	2×Intel 82574L 千兆比以太网	2×Intel 82573E 千兆比以太网
	电源模块	720W (1+1 备份）	520W
操作系统		ClientOS release 5.5 Final (Linux Kernel:2.6.18-194.el5 x86-64)	
软件		JDK1.6.0-16,Hadoop 0.20.2, Mastiff-0.1.2,Hive-0.5.0	



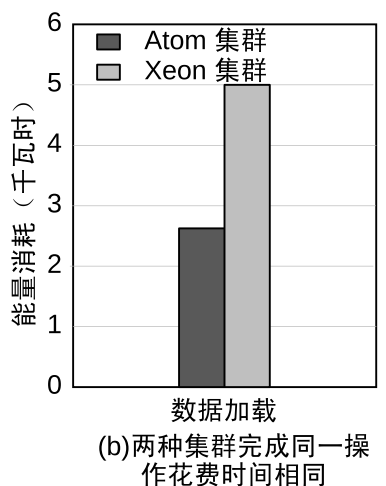


图1. Atom 和 Xeon 两种集群在(a)、(b)两种条件下能耗的比较[5]

1.2 微服务器具有更好的性价比

根据市场调查，互联网服务业是服务器市场中增长最快的部分，每年增长 40~65%。惠普实验室的凯文.林（Kevin Lim）等研究者开发了一个模仿数据中心互联网服务负载的基准

测试程序套件[6]。由互联网搜索（Web Search）、互联网邮件（Web Mail）、YouTube 和 MapReduce 组成，可以分别作为无结构数据、用户交互、多媒体和互联网平台（web as a platform）这四种互联网服务的代表。

凯文.林小组以总拥有成本（TCO，Total Cost of Ownership，该小组将其定义为硬件购置成本加上三年的电费）作为“价格”扩展了“性价比”的概念[6]，三年电费由所有部件功率的总和与时间的乘积来计算。由于服务器不会总是工作在满负荷功率[7]，因此以满负荷功率乘以 0.75 来估计平均功率。他们对各种成本的分析得出几点引人注意的结果：能源和硬件的花费是相当的；CPU 的购置成本和 CPU 的能耗是两项最大的花费；内存和磁盘的购置成本和能源花费也是相当大的；要想达到有竞争力的性价比，需要综合考虑系统各个部件的成本和能耗。

在传统的数据库或高性能服务器设计中，人们尽量提高单节点的性能。但是互联网服务的特性决定了性能的提高更多要依赖于服务器的数量。大型的互联网服务公司，如谷歌（Google）成功地利用了商用的低端服务器来建造他们的服务器。在这背后的逻辑是规模决定了成本，低端服务器巨大的市场规模极大降低了购置价格。

凯文.林小组对比了六种不同配置的服务器，以中端的传统服务器为基准评估了其它五种配置的新型服务器性价比。六种配置及其消耗功率和价格如表 2 所示:

表2. 六种不同级别的服务器配置[6]

系统	相当于	CPU	内存	功率(瓦)	价格($)
服务器 1	Xeon MP, Opteron MP	2p x 4 cores, 2.6 GHz, OoO, 64K/8MB L1/L2	FB-DIMM	340	3294
服务器 2	Xeon, Opteron	1p x 4 cores, 2.6 GHz, OoO, 64K/8MB L1/L2	FB-DIMM	215	1689
桌面	Core2, Athlon64	1p x 2 cores, 2.2 GHz, OoO, 32K/2MB L1/L2	DDR2	135	849
移动	Core2 Mobile, Turion	1p x 2 cores, 2.0 GHz, OoO, 32K/2MB L1/L2	DDR2	78	989
嵌入式 1	PA Semi, Emb.Athlon64	1p x 2 cores, 1.2 GHz, OoO, 32K/1MB L1/L2	DDR2	52	499
嵌入式 2	AMD Geode, VIA Eden-N	1p x 1 cores, 600MHz, inord.,32K/128K L1/L2	DDR1	35	379

“服务器 1”和“服务器 2”分别代表中端和低端服务器系统，“桌面”代表桌面系统，“移动”代表移动系统，“嵌入式 1”和“嵌入式 2”分别代表中端和低端的嵌入式系统。所有六种系统都配以 4GB 内存，服务器 1 配以 1.5Krpm 硬盘和 10G 以太网，其他系统配以 7200rpm 硬盘和 1G 以太网。

与基准配置服务器 1 相比，在单节点上看，其它所有系统的硬件成本有很明显的下降，最大的折扣来自于 CPU，DDR2 内存相比 FB-DIMM 也带来不少的节省。桌面配置的价格仅是服务器 1 配置的 25%，而嵌入式 1 则只有 15%。由于低功耗部件的额外成本，移动设备价格略高于桌面设备。功耗的下降与硬件成本的下降有着很相似的趋势。桌面系统与服务器 1 相比节省了 60%的能源，而嵌入式 1 则节省了更多的电力，达到了 85%。

表3. 四类应用程序中不同级别服务器的性能和性价比对比[6]

	负载类别	服务器 2	桌面	移动	嵌入式 1	嵌入式 2
性能	互联网搜索	68%	36%	34%	24%	11%
	互联网邮件	48%	19%	17%	11%	5%
	YouTube	97%	92%	95%	86%	24%
	mapred-wc [i]	93%	78%	72%	51%	12%
	mapred-wr [ii]	72%	70%	54%	48%	16%
	调和平均值	**71%**	**42%**	**38%**	**27%**	**10%**
性能/购置成本	互联网搜索	133%	139%	112%	175%	93%
	互联网邮件	95%	72%	55%	83%	44%
	YouTube	188%	358%	315%	629%	206%
	mapred-wc	181%	302%	241%	376%	101%
	mapred-wr	141%	272%	179%	350%	140%
	调和平均值	**139%**	**162%**	**125%**	**201%**	**91%**
性能/功耗	互联网搜索	107%	90%	147%	157%	103%
	互联网邮件	76%	47%	73%	75%	49%
	YouTube	152%	233%	413%	566%	229%
	mapred-wc	146%	197%	315%	338%	113%
	mapred-wr	114%	177%	235%	315%	157%
	调和平均值	**112%**	**105%**	**164%**	**181%**	**101%**
性能/总体拥有成本	互联网搜索	120%	113%	124%	167%	97%
	互联网邮件	86%	59%	62%	80%	46%
	ytube	171%	291%	351%	600%	215%
	mapred-wc	164%	246%	268%	359%	106%
	mapred-wr	128%	221%	200%	334%	147%
	调和平均值	**126%**	**132%**	**140%**	**192%**	**95%**

[i] 字数统计

[ii] 分布式文件写入

不出所料，低端的配置与服务器 1 相比性能有一定差距，不同系统在不同基准测试程序（Benchmark）下的性能差距各不相同。凯文.林小组给出如表 3 的实验结果，MapRduce 和 Youtube 与互联网搜索和互联网邮件相比，性能差距较小。这表明前两种负载不是 CPU 密集型的，性能主要决定于网络或磁盘。桌面系统在 MapRduce 和 Youtube 测试中性能下降了 10-30%，而在互联网搜索和互联网邮件测试中下降了 65-80%，嵌入式 1 则分别是 20-50% 和 75-90%。嵌入式 2 性能下降很明显，对于所有类型的负载，都有所下降。

综合考虑性能、价格和功耗，比较各配置下所得到的每瓦功耗的性能（performance/Watt）和每单位价格下的性能（performance/$）：与服务器 1 相比桌面、移动和嵌入式 1 都有很明显的优势，只有嵌入式 2 表现不佳。比如嵌入式 1 在 MapRduce 和 Youtube 中每单位总体拥有成本下的性能提高了 3~6 倍，在互联网搜索中也提高了 60%。除服务器 1 之外的所有系统由于 CPU 性能的明显劣势，每单位总体拥有成本下的性能在互联网邮件中有所下降，只有嵌入式 1 仍与服务器 1 相当，好于其它系统。

整体看来，比较低端的消费市场系统具有较高的性价比。谷歌的经验已经证实了桌面系

统的高效性，但是更值得注意的是嵌入式系统更有潜力，但系统的选择很重要（比如嵌入式 1 和 2 有很大差别）。另外按照服务器 1 的配置，每个 42U 机架消耗 13.6KW 电力，而嵌入式 1 仅消耗 2.7KW 电力，因此散热系统可以更简化，密度可以提高，进而进一步节省成本。

1.3 微服务器具备更好的集成密度

机柜设计方面：微服务器的处理器因为功耗较低，通常没有风扇，芯片管脚数目、每个处理器所带内存数目也较高端服务器有明显减少。这种特点给系统设计带来了新的挑战和机遇。一方面，人们有机会在同样的空间集成更多的处理器芯片，例如，SeaMicro 公司在将 384 个双核 Atom（或者 64 个 Xeon）处理器、64 个硬盘、64 个千兆网口（或者 16 个万兆网口）集成到一个 10U 的机箱里[4]，惠普将 288 个 EnergyCore（ARM Cortex A9）集成到 4U 的机箱里[8]。另一方面，在高密度集成的环境里面，设计者必须要考虑如何重新设计机柜的散热系统。根据凯文.林小组的分析，除处理器购置之外的第二大费用来自于能源与散热。低功耗系统使用更小规格的电路板，密度更大，为了取得足够的散热效果并且减少散热成本，在散热方面需要做更多的优化。

综上，微服务器在功耗、价格和体积上较传统的服务器都有很大的优势。本文的目的是促进科研人员和工业从业人员一起从系统和芯片两个角度讨论系统结构的设计和实现，并探索和分析运行在微服务器系统上的应用程序。本文后继将会分别介绍工业界的产品和学术界的研究前沿。

2 工业界的产品

2.1 SeaMicroSM10000

2.1.1 概览

SeaMicro SM10000 是一个服务器家族[4]，计算、存储、交换、管理和负载平衡等所有资源都集中在一个系统里。家族成员有三个：SM10000-64、SM10000-64HD 和 SM10000-XE。它们在一个 10U 高的机箱里分别集成了 256 个 Atom、384 个 Atom 或 64 个 Xeon 低功耗处理器。与市场上最好的集群服务器产品相比，SM10000 只需要 1/4 的电力和 1/6 的空间。

所有 SM10000 系列的计算机都是 10U 高、30 英吋长，是 x86-64 架构，支持即插即用，并且运行现有的操作系统、应用程序、管理工具，不需要第三方驱动程序，也不需要更改或重新编译任何软件。

所有 SM10000 结构相同，都包含 64 个计算卡、8 个存储卡和 8 个网络卡。所有这些卡通过机箱内一个高带宽低延迟的网络连接在一起，这个网络称为 Freedom™ Supercomputer Fabric 可以提供 1.28Tb/s 的带宽。

2.1.2 计算卡

SM10000 家族成员之间的不同只体现在计算卡，其它子系统都相同。SM10000-64 和 SM10000-64HD 使用的是 Atom 处理器，SM10000-XE 使用的是 Sandy Bridge Xeon 处理器。SM10000-64 计算卡上包含 4 个双核四线程 1.66GHz x86-64 Atom CPU 芯片，每个 CPU 芯片配以 4GB DDR3 内存。64HD 与 64 的区别仅是 CPU 芯片的数量增加到 6 个，内存仍然是每芯片 4GB，64HD 计算卡有 6 个 Atom 处理器，RAM 芯片分布在其周围（包括线路板背面），有 4 个 SeaMicro Freedom ™ ASIC（专用集成电路）通信芯片。每个 XE 计算卡上包括一个四核八线程 2.4GH Xeon 处理器，并配以 32GB DDR3 内存，线路板两面各有两个内存条，

通信芯片与 64 和 64HD 同样配置。

2.1.3 存储卡

每个存储卡上可以安装八个支持热插拔的 2.5 英寸 SATA 硬盘或固态硬盘。硬盘驱动器可以绑定在一起组成磁盘阵列（RAID）。不同于以前的服务器，磁盘绑定于一个节点，SeaMicro 存储结构具有很强的弹性，可以更高效地利用磁盘空间。磁盘空间可以被分割成薄片（slices），称为虚拟磁盘。虚拟磁盘可以分配给任何节点，也可以在多个节点间只读共享。

2.1.4 网络卡

每个网络卡可以支持 8 个千兆或两个万兆以太网接口。装满 8 个网络卡的系统可以设置为 64 个千兆接口或 16 个万兆接口，这些接口相当于架顶交换机（top-of-rack switch）的上行接口，用于连接行交换机。万兆网络使用的是 SFP+接口，可以支持光纤和双绞线。

2.1.5 总结

SM10000 是一个标准的 x86-64 服务器系列，可以作为由 60 个 1U、双网口、四核服务器节点，架顶交换机、终端服务器和负载平衡器组成的服务器系统的替代品。SM10000 使用 Intel Atom 或 Xeon 处理器，只需市场上同类产品 1/4 的耗电量，占用 1/6 的空间，同时不需要对软件做任何改动。

2.2 HP Redstone

2.2.1 HP Moonshot 计划

惠普在 2011 年 11 月宣布了 Moonshot 的计划[8]，开发用于企业数据中心的高能效服务器。此计划致力于在数千服务器节点之间共享存储、网络、管理和散热资源，从而实现超大规模（hyperscale）计算环境的搭建。Redstone 是此计划设计的第一个服务器平台，是利用德州创业企业 Calxeda 的 ARM 处理器 EnergyCore 开发的服务器产品。

2.2.2 Redstone 的搭建

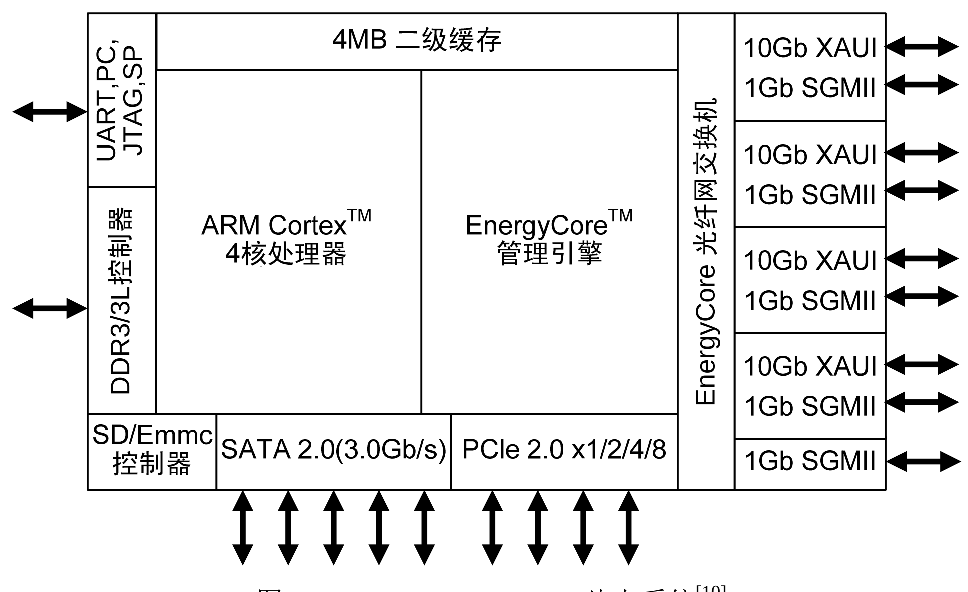


图2. Calxeda EnergyCore 片上系统[10]

EnergyCore 是专门为超大规模服务器负载设计的 ARM 处理器[9]。第一款 EnergyCore 产品 ECX1000 是 32 位双核或四核处理器。芯片上集成了内存、读写（I/O）、和存储控制器，还有一个 2 层交换结构，如图 2 所示。这个交换结构称为 EnergyCore Fabric Switch，可以支持二维环面（2D torus）、网状网（mesh）、胖树（fat tree）、蝶形树（butterfly tree）等网络拓扑。

Calxeda 将 4 个 ECX-1000 芯片和 4 个内存槽放在一个板卡上组成一个 4 节点的服务器，称为 EnergyCard，如图 3，上部是四个 DDR3 内存插槽，每个芯片旁边三个和最右侧的四个是 SATA 接口，连接到芯片内的 SATA 控制器。最下方的 PCI[1]连接器可以插到背板（passive backplane）上，从而取得电源并与其它板卡的互联通路。由于内存控制器、网络部件和其它读写部件都和 CPU 核一起集成在一个芯片上，EnergyCard 是一个很精简的小板。Calxeda 在 2011 年初完成的最早的参考设计是在一个 2U 的机箱里放入了 120 个芯片，也就是 120 个节点。密度达到 60 节点/U。



图3.　Calxeda 4 节点服务器板卡[11]

惠普使用了高度为 2U 的半宽 ProLiant 托盘和 ProLiant SL6500[12]机箱搭建 Redstone 服务器。托盘里插入了三排，每排六个 EnergyCard。这样每个托盘里就放置了 72 个服务器节点。一个 SL6500 机箱里可以插入 4 个托盘，共 288 个节点。密度达到了每 U 空间 72 个节点。与 Calxeda 的参考设计相比，密度增加了 20%。

Redstone 使用的 SL6500 机箱里有三个互为备份的电源供应模块，组成一个电源池。即使其中一个模块发生故障，整个系统也可继续工作。散热风扇的数量是八个。每个托盘前面有四个 10Gb/s 的数据接口，这四个接口连接到内部的 EnergyCore Fabric Switch。所有这些接口可以通过 10G 以太网接口（XAUI）互联，最多可以将 4096 个节点连接在一起。（当前 4000 节点对于 Hadoop 集群已经是一个很大的规模）

虽然可以通过托盘前面的四个接口互联，但惠普推荐的接法是一个 SL6500 机箱内部的四个托盘用集成的互联结构连接，而多个机箱之间则通过两条 10Gb 电缆连接到机架顶部的交换机上。这样的连接方法相当于把 SL6500 看作一个带有架顶交换机的机架，而外部的 10G 交换机则扮演了列尾交换机（end-of-row switch）的角色，将多个机架连接在一起。

在 Redstone 系统中，默认将磁盘放置在外部的磁盘阵列，访问磁盘要通过网络。但是也可以牺牲节点的密度，腾出空间，通过 SATA 接口在 EnergyCard 上插入固态硬盘或 2.5 寸硬盘。每个托盘内最多可以插入 192 个固态硬盘或 96 个硬盘。这些硬盘或固态硬盘可以从背板获取电源。

2.2.3 Redstone 的现状

[1] Peripheral Component Interconnect

目前，半个机架可以容纳 1600 个 Redstone 服务器节点和它们使用的交换机。整个系统共使用 41 条通信电缆，总功率为 9.9 千瓦，价格为 120 万美元。虽然一个传统的处理能力相当的 x86 集群，仅需要 400 个 Xeon 节点，但这 400 个节点却需要 10 个机架、1600 条通信电缆和 91 千瓦的电力，价格为 330 万美元[13]。

当然 Redstone 与传统服务器是有区别的，其时钟频率只有 1.1 或 1.4GHz，位宽只有 32 位，单节点内存只有 4GB，只有部分负载可以高效地利用硬件资源。惠普认为网页服务和海量数据处理在这种环境下应该会有不错的表现。

惠普正在从德克萨斯的休斯顿（惠普的 PC 和服务器工厂所在地）开始，逐步把 Redstone 服务器安装在世界各地的探索实验室（DiscoveryLabs），并且让潜在的客户把他们的程序上传到 Redstone 服务器上在 Canonical Ubuntu 或 Red Hat Fedora Linux 上运行。

3 学术界的前沿

3.1 Nanostore

世界上的数据正在以爆炸的形式增长，速度远超摩尔定律。比如谷歌索引的在线数据 2002 年是 5EB[2]，到 2009 年增长到 280EB，7 年增长了 56 倍[14]。而摩尔定律在这 7 年里只能给计算机性能带来 16 倍的增长。最近的一项估计显示，每分钟有 24 小时的视频上传到 YouTube。以 2-5Mbps 的码率计算，每天将产生 45-75TB 的数据。更近一些，大规模的传感器部署也加剧了数据的爆炸速率。纳米级传感器的发展使人们可以实时、细粒度地采集多种数据，包括震动、倾斜、旋转、气流、光、温度、化学信号、湿度、地理位置等等。看到了这些传感器技术的发展，科研人员计划开发一个“地球中枢神经系统”(CeNSE)[15]。这个系统将利用广泛分布的传感器网络在很多领域发挥有趣的作用，可以深入零售、保安、交通、地震、石油勘探、天气和气候、野生动物跟踪等各个方面。但是这个美丽的前景将带来前所未有的数据量和数据处理负载。

移动电子设备在世界范围内的普及程度持续上升。这些电子设备具有收集和发布信息的能力。它们在不停地产生实时的丰富的数据。比如在迈克尔.杰克逊过世的 2009 年 6 月，据估计每分钟有 5000 条微博发布到 Twitter，而 AT&T 则每分钟为用户传送 65000 条短信。在一个 90 天的时间段里，20%的网络搜索访问的是典型的“新数据”[14]。值得注意的是新的数据具有很高的多样性，它们可以是文字、音频、视频、图像……中的任意一种或多种的组合。数据具有多样性的同时，对这些数据的组织方式也多种多样，包括有结构的存储（可以通过数据库访问）、无结构存储（以文件的形式保存）或者半结构化的存储（如 XML、e-mail）。

数据的增长推动了以数据处理为中心的应用程序的发展。应用程序对数据的操作多种多样，如：捕捉、分类、分析、处理、存档等等。这些操作的应用实例更是数不胜数，如网页搜索、推荐系统、决策支持、在线游戏、排序、压缩、传感器网络、特殊查询、多维数据服务（cubing）、多媒体转码、流媒体、照片处理、社交网络分析、个性化、自动摘要（summarization）、索引建立、歌曲识别、聚合（aggregation）、混搭（Web mashups）、数据挖掘，还有加解密等等。

与事务处理和网页服务这样的传统负载相比，新出现的工作负载是以数据为中心的，它们使系统设计中的很多假设发生了改变。新型的负载数据规模更大，操作更多样，更复杂。

[2] 10^{18}（百万万亿）字节

数据负载增长的同时，技术方面也有一些新的趋势。计算方面，最近的微处理器倾向于多核设计，强调多个简单的核心，以获得更大的吞吐量；处理器核芯片的工作电压越来越接近临界值，以提高性能功耗比[16]。网络方面，为了满足大量不同计算单元之间的通信，带宽大大增加。

然而，与数据中心相关的最重要的技术方面的变化是非易失存储器的发展和使用。闪存（Flash）已经广泛地应用到前沿消费市场，如 iPhone，在企业市场也有一定的应用，如 Fusion-io。图 4 展示了当今常用存储器成本的变化趋势。这个趋势表明非易失存储器是 DRAM 的一个可能的替代物。与传统的硬盘和 DRAM 对比，新出现的非易失存储器件显示出与闪存相似的特性，值得注意的有相变存储器（phase-change memory，PCM）[17]和更新出现的忆阻器（memristors）[18]。未来的非易失存储器可能成为 DRAM 的替代物，可以达到与 DRAM 相当的速度、更低的功耗，具有磁盘一样的非易失性而不需要磁盘马达消耗的额外电力。最近的研究情况显示，DRAM 容量的增长速度有所下降[19]，而非易失性随机访问存储器（NVRAM）有潜力取而代之。

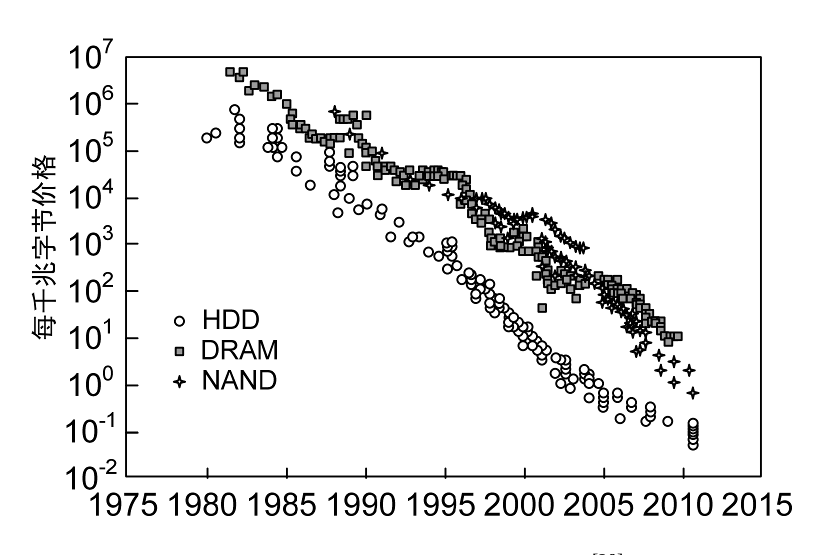


图4. 常用存储器成本的变化趋势[20]

密度和寿命是 NVRAM 的两大技术问题，但是最近的趋势显示这些问题可以解决。多层设计可以增大密度[21]，在单芯片范围内，使用穿透硅通孔（through-silicon vias）互连的三维层叠芯片可以得到很高的密度。这样的三维层叠的另一个好处是方便把处理器和存储器集成在一起，这样可以得到更高的带宽和更低的功耗。寿命方面，与闪存相比相变存储器和忆阻器表现得更好，每个单元可以写入 10^7~10^8 次，闪存则仅 10^5 次。只是这两种器件技术的成熟和扩大应用还需假以时日。更多的关于非易失存储器的信息可以在近期的一些综述和专题报道中找到[22][23]，如 HotChips 2010。

以上这些趋势显示，把相变存储器、忆阻器这些技术，尤其是与三维层叠、多核和先进的网络互联放在一起思考，会引起更根本的系统结构的创新，而不是仅利用它们构建层次存储结构中新的一层，或者仅制造固态硬盘。

这为重新思考计算机系统结构和内存层次提供了一个不可多得的机会。Nanostore 这个名词的提出综合了当前微处理器反映出的纳米技术和以数据为本代替以计算为本的思想[20]。Nanostore 的主要特点是把微处理器和非易失存储器集成在一起，从而去除中间的很多存储结构层次。所有的数据都保存在单个层次的非易失存储器里，而不再使用传统的磁盘和 DRAM 这两层结构，磁盘可以从系统中移出去，作为备份装置。

比如说一个 Nanostore 芯片可以由一个三维堆叠的高密度的非易失存储器（相变存储器或忆阻器）和一个顶层的高能效的计算核心组成。计算核心与非易失存储器之间以穿透硅通孔互连，以取得高带宽低延迟的数据通信。每个 Nanostore 芯片都配以网络接口，都是一个五脏俱全的系统。很多个芯片通过板上的网络互联可以以任何一种拓扑结构（古老的胖树或新出现的 HyperX[24]等等）连接在一起，很多 Nanostore 芯片相连就组成了一个大规模的分布式系统，与现有的用于数据本位型（data-centric）计算的大规模集群极为相似。

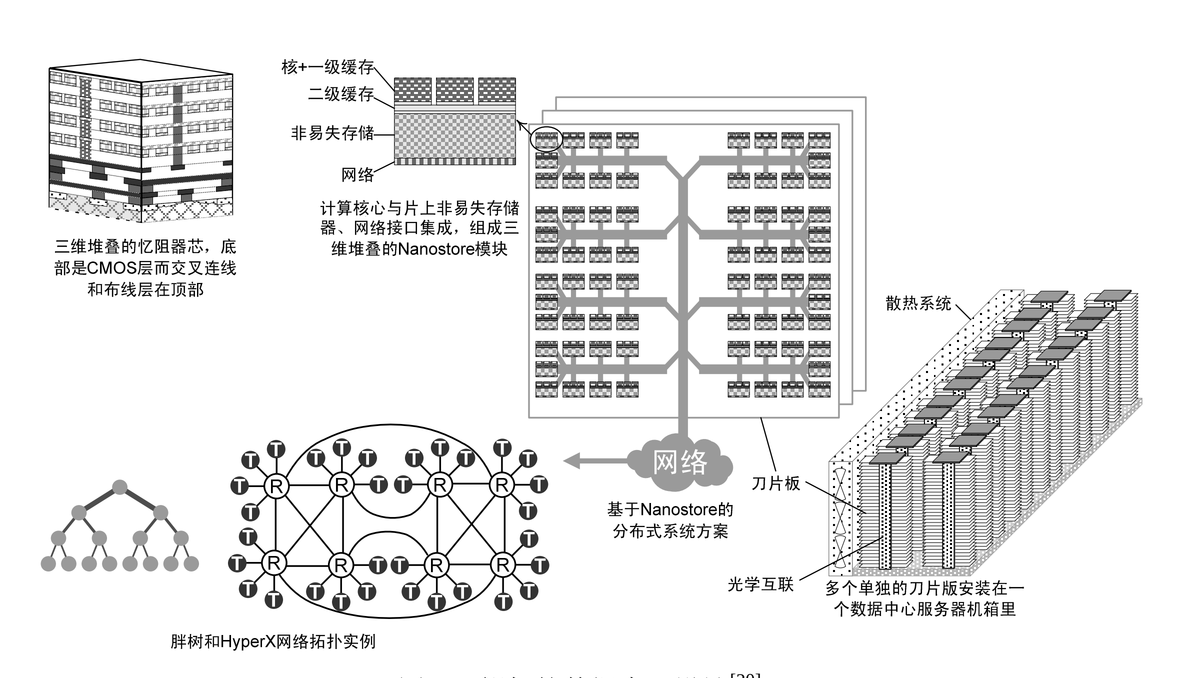


图5.　假想的数据中心设计[20]

Nanostores 将处理器核心和非易失存储置于同一芯片并将芯片相互连接组成一个更大的集群来处理以数据为中心的负载流

在更高层面，很多 Nanostore 芯片可以集成在一个小的子板（micro blades）上，然后把这些小板子插到一个经典的刀片服务器背板上。假设这些子板的散热特性已经确定，可以预想一个新的板级组装技术。如图 5，是一个假设的数据中心设计，很多刀片服务器以优化的散热和很高的密度连接到一个光学的背板上。

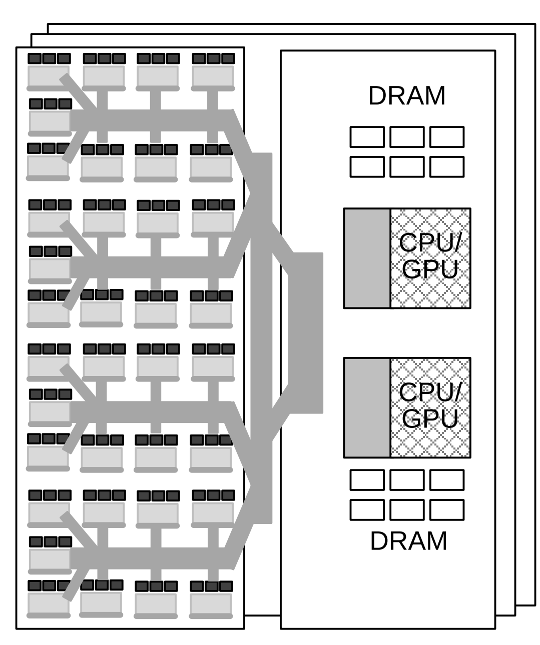


图6.　在一个层次化计算体系中外加的计算能力可以和 nanostore 结合

供电和发热是三维层叠芯片的一个重要问题，它限制了 Nanostore 芯片中计算单元的数量。图 6 展示了如何加入额外的更强大的计算单元，来支援芯片上的计算单元，这就构建了一个多层次的计算系统。这样 Nanostore 系统可以像现在常见的系统一样具有强大的计算单元和多层的存储结构，可以处理计算密集型负载，也可以处理遗留到未来的旧式的工作负载。

计算单元放在非易失存储单元的附近，基于这个前提可以有多种不同的设计。如图 7 所示，这三种方案很好地对比了不同设计的得失。第一种是较传统的使用 DRAM 和固态硬盘的设计，第二种是 side-stacked （旁侧堆叠）Nanostore，第三种是三维堆叠（3D-stacked）Nanostore。第三种设计与之前的假想设计很相似，计算单元堆叠于存储单元顶部，二者通过通孔相连。而第二种结构是 Nanostore 的另一种方案，其中计算单元与存储单元距离很近却是分离的，二者彼此相邻，通过总线相连。

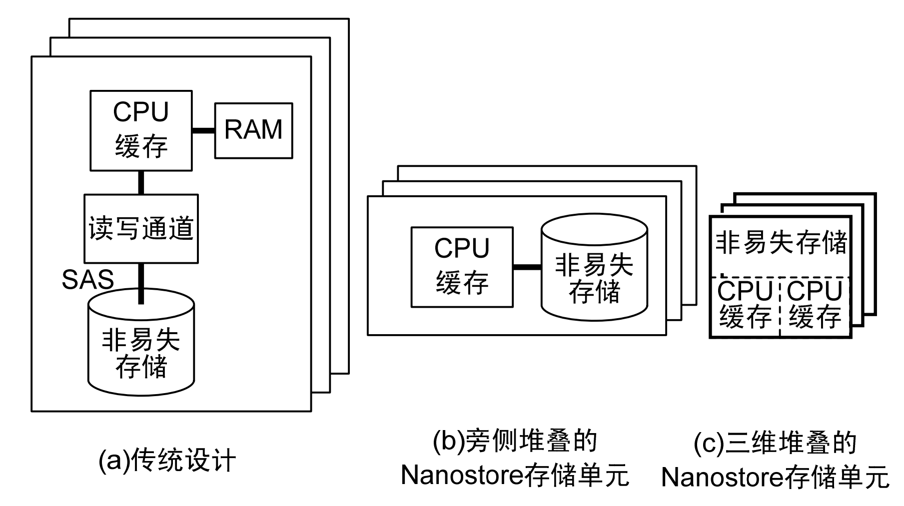


图7. 三种设计为数据为中心的工作流负载处理提供不同折中选择[20]

从数据本位的工作负载的角度看，需要关心的问题是单位数量的数据可以得到多少计算资源和影响这些资源被有效使用的瓶颈在何处。

传统的设计比较适合于计算繁重，而通信带宽较小的负载（比如视频转码），或者是数据集中，热数据和冷数据在数量上相差多个数量级的负载（比如图像存储）。在这种结构中，需要更多的带宽以访问更下层数据；而要更好地利用数据并行性的负载（比如 MapReduce、排序、点击流和日志分析）则三维堆叠和旁侧堆叠 Nanostore 可能更适用。

Nanostore 数据延迟较小，重写软件以利用这个特性可以带来额外的好处，但最终会遇到网络连接的瓶颈。对于高并行性的负载，我们需要按照更小的粒度并行化，实现较少的跨节点通信，这种情况下 3D-stacked Nanostore 会工作得最好，但简单的计算单元会成为瓶颈。虽然实现高效的并行化需要一定的工作量，但是在开销、成本和能耗方面的优势会证明这些努力是值得的。

数据本位负载对带宽和延迟的需求、软件的并行化进程以及本地网络互联的改进都是未来的计算机系统采用 Nanostore 设计的支持因素。

3.2 新的存储

为了处理越来越多的数据，服务器的设计从提高单节点性能（纵向扩展，scale-up）向增加节点数目（横向扩展，scale-out）转变[25]。为了提高密度、提高性能功耗比、降低总体拥有成本，横向扩展体系结构使用大量中等性能 CPU，而不是少量高性能 CPU，且输入输出（I/O）资源一般会与计算资源分离，并在计算资源之间共享。在分离式的结构下，计算资源和读写资源可以独立地增加或减少。

前人开发和使用了很多存储共享系统方案，比如 iSCSI（Internet Small Computer Systems Interface）和 FC（FibreChannel）[26]。虽然这些方案已经广泛地在工业界使用，但它们都是针对存储域网（SAN，Storage area network）的[26]，并不适合高密度的微服务器，因为微服

务器一般每个机箱里集中了几十到几百个 CPU，并且会跟几十个高性能的存储设备相连。因此，为了适应微服务器的需求，英特尔的廖广登（Guangdeng Liao，音译）等人提出了一个新的高效的基于块的存储系统 Light Peak Block Transport 或者叫做 LBLK[27]。这不仅是在微服务器之间共享分离式存储资源的一个方案，也是一个促进微服务器研发的关键平台技术。

LightPeak 也称为 Thunderbolt，是英特尔开发的一个光学连接技术，用于计算机和消费电子产品之间的连接。这个技术已经用到了苹果的 MacBook Pro 笔记本电脑。LightPeak 通过光纤提供 10Gbps 或更高的带宽，并且由于使用了硅光子技术（Silicon Photonics）技术，成本极低。

LightPeak 可以同时传送多种 I/O（输入输出）协议，并且以较小的数据包获得了很低的延迟。LightPeak 还提供了基于优先级的带宽分配和回收机制。

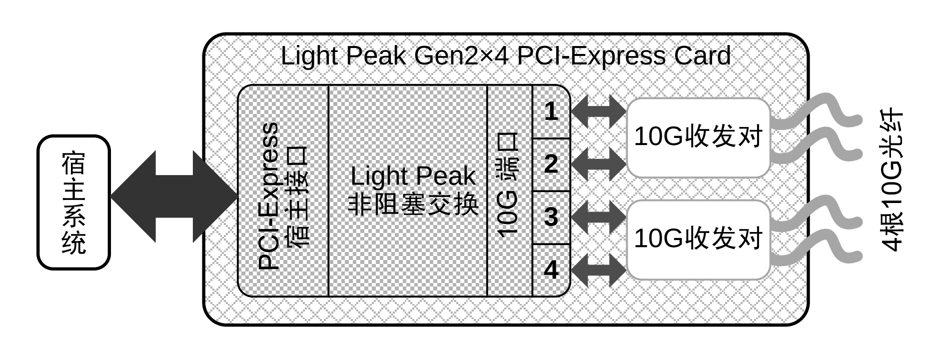


图8. LightPeak PCI-E 卡的框图[27]

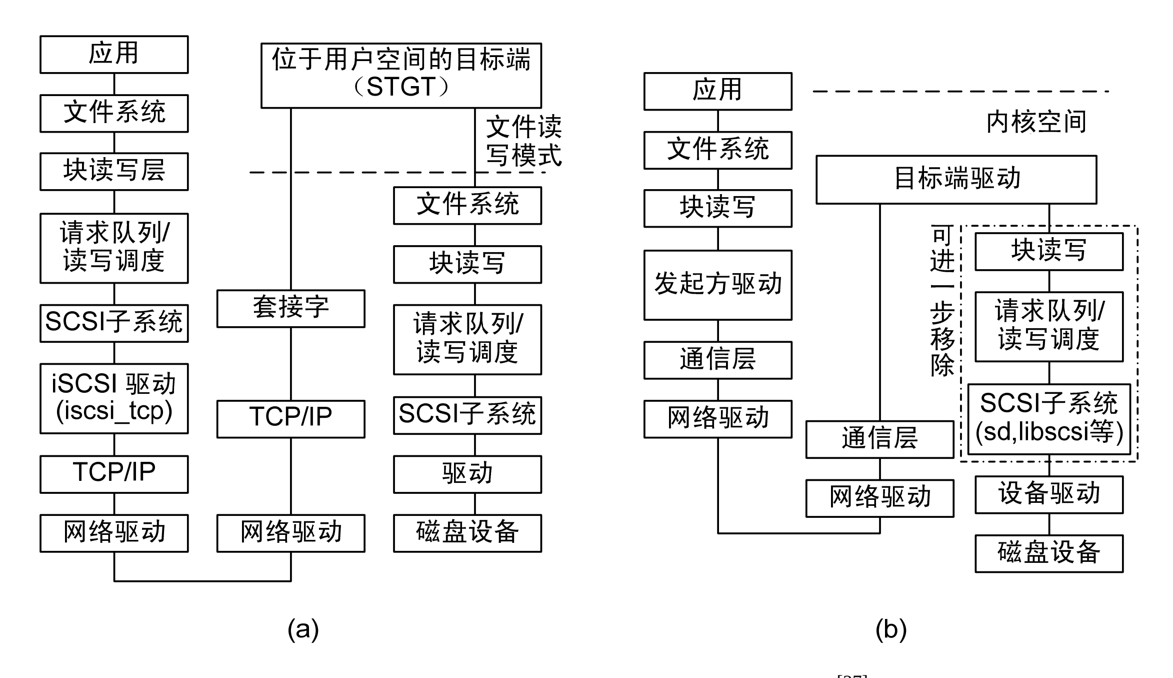


图9. (a)iSCSI 执行流程；(b)LBLK 执行流程[27]

廖等人使用 LightPeak 网卡来搭建 LBLK 平台[27][28]。如图 8，这是一种 PCI-E 接口的扩展卡，由一个内部的非阻塞交换机和配以两个光学收发器对的四个 10GB 接口组成。每个收发器对连接两个 10GB 光纤，并负责光电转换。内部的交换机可以同时处理 40Gbps 的接收任务和发送任务，而且可以在不影响 CPU 的情况下直接把数据从一个网口传送到另一个。鉴于这些优点，微软也在做关于 LightPeak 的实验，以期用于数据中心的网络连接。

与传统的 iSCSI 存储系统类似，LBLK 系统也是由发起方（initiator）和目标（target）组成，应用程序在发起方进行文件操作，目标服务器接收发起方发来的命令，执行后返回结果。

LBLK 的目标端是一个基于 Xeon 的节点。这个“存储节点”对外抽象出两个视角，存储块（Storage Blocks）和磁盘分区（Storage Partitions）。存储块是一个固定大小的数据块，是传输中的最小单位；磁盘分区是一段连续的存储块，里面的存储块从零开始编号。一个分区只能分配给一个服务器节点，一个服务器节点可以得到多个分区。在英特尔的实现中，为每个节点分配四个分区 boot、file system、swap 和 data。

虽然与 iSCSI 类似，但如图 9 所示，LBLK 缩短了执行流程。发起方驱动直接与 Linux 块读写层通信，并把块读写（BIO，Block I/O）请求转换成通信层（communication layer）的数据流。这个方案跳过了 SCSI 层和基于请求队列的读写调度。跳过读写调度的理由是，微服务器的 CPU 功能较弱，越来越流行的固态硬盘不需要这个占用 CPU 时间的读写调度器。

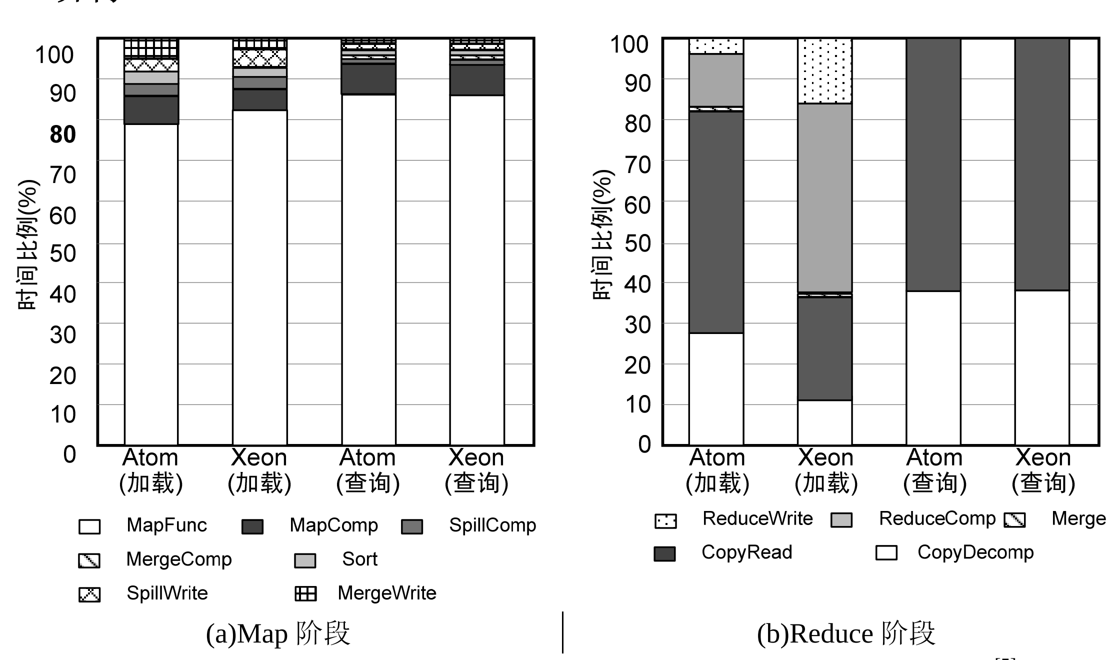
通信层是英特尔专门为微服务器环境定制的，用于机箱之间的通信。在通信层多个数据流通过多个物理通道顺序传输。通信层使用基于余额（credit）的流量控制算法[29]，设计和实现相比 TCP/IP 要简单很多。

在目标端，驱动程序是一个内核态的多线程程序，这样就避免了开销昂贵的用户/内核上下文切换。驱动程序把接收到的命令转换为块读写请求。虽然当前 SCSI 和请求队列还在使用，但英特尔计划使用一个与 NVM Express 类似的基于块读写的驱动程序替换它们。

为了测试性能，廖等人将 LBLK 部署在由 4 台 Atom 服务器和一台“存储节点”组成的集群中，同时也在 LightPeak 上部署了 iSCSI，用于对比。结果显示 LBLK 占用的 CPU 时间比 iSCSI 更少，同时提供的带宽更大。

4　未来的发展方向

4.1 异构



(a)Map 阶段　(b)Reduce 阶段

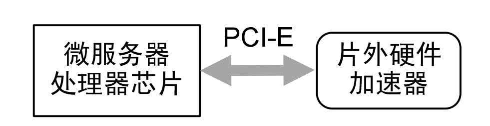
图10.　Xeon 和 Atom 两种不同的计算机集群对不同操作 CPU 时间分解比较[5]

由于微服务器单线程的处理能力不强，在其发展过程中需要考虑如何适应更多的应用。从系统结构的角度来讲，需要设计异构系统，即将微服务器和专用的加速器或者更强大的高端处理器有机结合起来。本节以专用加速器为例进行阐述。

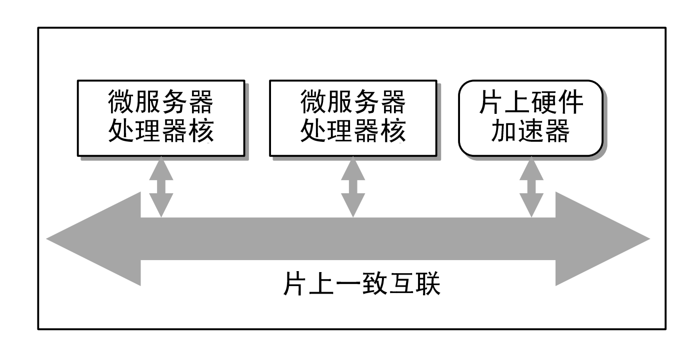
前文中提到我们在 Xeon 和 Atom 两种不同的计算机集群上测试 Mastiff 的实验不仅验证了低功耗处理器的高性能功耗比，通过对工作负载的时间分解，我们还可以看到某些固定的操作过程（压缩和解压）占用了不小比例的 CPU 时间。图 10(a)显示了对 Map 阶段的时间分解，可以看出虽然 Map 任务本身占用了绝大部分 CPU 时间，但是压缩操作也占有不小的比例，Atom 服务器上占用 11%，Xeon 服务器上占用 7%。图 10(b)是 Reduce 阶段的时间分解。在 Reduce 阶段大部分时间被 Hadoop 框架占用而不是 Reduce 任务本身。在所有的测试实例中压缩和解压占用了相当大的时间比例。比如在 Atom 集群上，数据加载（data load）和数据查询（data query）任务中压缩和解压占用的 CPU 时间比例分别是 41.2%和 37.9%。这是相当大的一个比例。

由于压缩解压占用了很大的一部分 CPU 时间，异构系统的提出成为很自然的事情。我们可以将压缩解压操作从 CPU 上分离出来，交给硬件加速器来完成。硬件加速器可以消除通用处理器上的时间开销，从而取得高出好几个数量级的加速比和性能功耗比。

实现以上这样的异构系统，图 11 给出了两个可能的选择：一是 PCI-E 加速卡，二是芯片集成加速器。前者适用于升级现有系统，而后者是更加根本的增强方式，需要重新设计芯片。



(a)片外实现方案



(b)片上实现方案

图11. 硬件加速器的实现[5]

4.2 分离

微服务器的研究范围需要从处理器向系统扩展，包括互联技术、存储技术，以及相关硬件加速工作。在设计中应该想方设法把可以从 CPU 上分离的负载分离出来。网络互联中的 TCP/IP 协议栈的处理可以由网卡上的高效的专用集成电路完成，从而节省弱 CPU 的时间。这样，无论是时间效率还是功耗效率都会有所提高。而将存储设备与计算资源分离开，则可以降低磁盘容量的冗余，无论从硬件成本还是功耗成本都会有所节省。

参考文献：

[1] “DEll PowerEdge.” [Online]. Available: http://www.dell.com/poweredge.

[2] “HP Proliant Servers.” [Online]. Available: http://h18004.www1.hp.com/products/servers/platforms/.

[3] “HP Shapes the Future of Extreme Low-energy Server Technology.” [Online]. Available: http://www.hp.com/hpinfo/newsroom/press/2011/111101xa.html?mtxs=rss-corp-news.

[4] “SeaMicro.” [Online]. Available: http://www.seamicro.com/.

[5] Xiaoyan Gu, Rui Hou, Ke Zhang, Lixin Zhang, and Weiping Wang, "Application-driven Energy-efficient Architecture Explorations for Big Data," in *First Workshop on Architectures and Systems for Big Data*, 2011.

[6] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, and S. Reinhardt, "Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments," *SIGARCH Comput. Archit. News*, vol. 36, no. 3, pp. 315–326, Jun. 2008.

[7] Chandrakant Patel and Parthasarathy Ranganathan, "Enterprise Power and Cooling: A Chip-to-DataCenter Perspective," presented at the Hot Chips 19, 2007.

[8] "Become a part of HP Project Moonshot." [Online]. Available: http://h17007.www1.hp.com/us/en/iss/110111.aspx.

[9] "Calxeda." [Online]. Available: http://www.calxeda.com/.

[10] "Calxeda ECX-1000 Key Features." [Online]. Available: http://www.calxeda.com/technology/products/processors/ecx-1000-features/.

[11] "Quad-Node EnergyCard." [Online]. Available: http://www.calxeda.com/technology/products/energycards/quadnode/.

[12] "HP ProLiant SL6500 Scalable System Family data sheet." [Online]. Available: http://www.hp.com/hpinfo/newsroom/press_kits/2010/HPOptimizesAppDelivery/SL6500_Scalable_System.pdf.

[13] "HP Project Moonshot hurls ARM servers into the heavens." [Online]. Available: http://www.theregister.co.uk/2011/11/01/hp_redstone_calxeda_servers/.

[14] Marissa Mayer, "Innovation at Google: the physics of data." [Online]. Available: http://www.parc.com/event/936/innovation-at-google.html.

[15] R. Stanley Williams, "A Central Nervous System for the Earth," *Harvard Business Review*, vol. 87, no. 2,2009, p. 39, 2009.

[16] B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Proceedings of the 2007 international symposium on Low power electronics and design*, New York, NY, USA, 2007, pp. 32–37.

[17] C. Lam, "Cell Design Considerations for Phase Change Memory as a Universal Memory," in *VLSI Technology, Systems and Applications, 2008. VLSI-TSA 2008. International Symposium on*, 2008, pp. 132 –133.

[18] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, May 2008.

[19] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," *SIGARCH Comput. Archit. News*, vol. 37, no. 3, pp. 2–13, Jun. 2009.

[20] P. Ranganathan, "From Microprocessors to Nanostores: Rethinking Data-Centric Systems," *Computer*, vol. 44, no. 1, pp. 39–48, Jan. 2011.

[21] D. L. Lewis and H. S. Lee, "Architectural evaluation of 3D stacked RRAM caches," in *IEEE International Conference on 3D System Integration*, 2009, pp. 1–4.

[22] H. Li and Y. Chen, "An overview of non-volatile memory technology and the implication for tools and architectures," in *Proceedings of the Conference on Design, Automation and Test in Europe*, 3001 Leuven, Belgium, Belgium, 2009, pp. 731–736.

[23] N. Jouppi and Y. Xie, "Emerging technologies and their impact on system design," in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, New York, NY, USA, 2009, pp. 427–428.

[24] J. H. Ahn, N. L. Binkert, A. Davis, M. McLaren, and R. S. Schreiber, "HyperX: topology, routing, and packaging of efficient large-scale networks.," in *SC*, 2009.

[25] D. G. Andersen, J. Franklin, A. Phanishayee, L. Tan, and V. Vasudevan, *FAWN: A Fast Array of Wimpy Nodes*. 2008.

[26] V. V. Riabov, "Storage Area Networks (SANs)," in *The Internet Encyclopedia*, H. Bidgoli, Ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2004.

[27] G. Liao, S. McGowan, T. L. Willke, J. Howard, and P. Cayton, "Efficient Block-level Sharing of

Disaggregated Storage for Microservers," in *the first Workshop on Architecture and Application Exploration of Micro-Server Systems*.

[28] S. Addagatla, M. Shaw, S. Sinha, P. Chandra, A. S. Varde, and M. Grinkrug, "Direct Network Prototype Leveraging Light Peak Technology," in *Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects*, Washington, DC, USA, 2010, pp. 109–112.

[29] B. C. Kung, T. Blackwell, K. Chang, H. T. Kung, and D. Lin, "Credit-Based Flow Control for ATM Networks," in *IEEE Network*, 1995, pp. 101–114.

作者简介：

王　聪：　中国科学院计算技术研究所，2011 级博士生，wangcong@ict.ac.cn
侯　锐：　中国科学院计算技术研究所，副研究员，hourui@ict.ac.cn
张立新：　中国科学院计算技术研究所，研究员 ，zhanglixin@ict.ac.cn

（ 上接第 41 页 ）

[19] OPNET Technologies Inc. OPNET modeler website. http://www.opnet.com/solutions/network_rd/modeler.html

[20] R. Buyya and M. Murshed. "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing." Concurrency and Computation: Practice and Experience, 14(13-15), Wiley Press, Nov.-Dec., 2002.

[21] Gartner Group, available at: http://www.gartner.com/

[22] The network simulator ns-2. http://www.isi.edu/nsnam/ns/

[23] R. Fujimoto. Parallel Discrete Event Simulation. Communications of the ACM, October 1990.

[24] Iperf: http://sourceforge.net/projects/iperf/

[25] IMB: http://software.intel.com/en-us/articles/intel-mpi-benchmarks/

作者简介：

胡农达：　中国科学院计算技术研究所，先进计算机系统实验室，博士研究生, hunongda@ict.ac.cn
付斌章：　中国科学院计算技术研究所，先进计算机系统实验室，助理研究员, fubinzhang@ict.ac.cn
隋秀峰：　中国科学院计算技术研究所，先进计算机系统实验室，助理研究员
李　龙：　中国科学院计算技术研究所，先进计算机系统实验室，硕士研究生
朱晓东：　中国科技大学，计算机系，博士研究生
李　涛：　美国弗罗里达大学，电子计算机系，副教授，博士生导师
陈明宇：　中国科学院计算技术研究所，先进计算机系统实验室，研究员，博士生导师, cmy@ict.ac.cn
张立新：　中国科学院计算技术研究所，先进计算机系统实验室，研究员，博士生导师